DOCUMENT RESUME

ED 037 787 CG 005 192

AUTHOR	Astin, Alexander W.; Boruch, Robert F.
TITLE	A "Link" System for Assuring Confidentiality of Research Data in Longitudinal Studies. ACE Research Reports, Vol. 5, No. 3.
INSTITUTION	American Council on Education, Washington, D.C. Office of Research.
PUB DATE	Feb 70
NOTE	22p.
AVAILABLE FROM	American Council on Education, Office of Research, One DuPont Circle, Washington, D.C. 20036
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.20
DESCRIPTORS	Computer Storage Devices, *Confidential Records, Data Bases, Data Processing, Filing, *Information Storage, *Information Systems, *Research Methodology, *Security, Tape Recordings

ABSTRACT

The purpose of this article is to describe a system for protecting the anonymity of subjects in longitudinal research and for maintaining the security of data files. The basic system comes from the American Council on Education Cooperative Institutional Research Program. Data is collected from questionnaires, and converted to magnetic tape. Original data is destroyed. Two separate tape files are then set up. The first file contains answers of the person, together with an arbitrary identification number. The second file has the person's name and address and the same arbitrary number. The former file is accessible to members of the research staff, the latter is locked in a vault. The "Link" system elaborates on the above scheme by removing identification numbers from the name and address file and substituting another unrelated number. A third file was then created, which contained only the two numbers. This file then links the subjects identity with his answers to questions. The link file is then deposited at a computer facility located in a foreign country. This file is released to no one. Follow-up data is then collected again from the same students, with the person's number used for identification. This information is then sent to the foreign center, where the second number is replaced by the first. The data is then merged with previous data for longitudinal research. (KJ)

OFFICE OF RESEARCH • AMERICAN COUNCIL ON EDUCATION

A "Link" System for Assuring Confidentiality of Research Data in Longitudinal Studies

ALEXANDER W. ASTIN

ROBERT F. BORUCH

VOL. 5 NO. 3 • 1970

ADDITIONAL COPIES OF THIS RESEARCH REPORT (VOL. 5, NO. 3, 1970) MAY BE OBTAINED FROM THE OFFICE OF RESEARCH, AMERICAN COUNCIL ON EDUCATION, ONE DUPONT CIRCLE, WASHINGTON, D.C. 20036.

A "Link" System for Assuring Confidentiality of Research Data in Longitudinal Studies

Alexander W. Astin

Robert F. Boruch

American Council on Education
Office of Research

ACE RESEARCH REPORTS
Vol. 5, No. 3
February, 1970

A "Link" System for Assuring Confidentiality of Research Data In Longitudinal Studies[1]

Alexander W. Astin
Robert F. Boruch
American Council on Education

Behavioral scientists have long recognized the importance of longitudinal research data in studies of human growth and development. A major logistical problem in such studies, however, is that the research subjects must be identified in some manner so that they may be resurveyed periodically. Although most researchers are aware that possessing identifying information imposes on them certain obligations to protect the anonymity of their subjects (Privacy and Behavioral Research, 1967), few attempts have been made to develop improved techniques for insuring data security and respondent anonymity. That such efforts are sorely needed is evident from massive anecdotal evidence (e.g., Westin, 1967), and from empirical and systemic studies (e.g., Nugent, 1969; Boruch, 1969).

The purpose of this article is to describe a system for protecting the anonymity of subjects in longitudinal research and for maintaining the security of data files. The system has been developed in connection with the Cooperative Institutional Research Program of the American Council on Education. However, we believe that its basic design is applicable, perhaps with only minor variations, to longitudinal research in education and other fields. One

of our major goals in bringing this system to the attention of the community of researchers is to encourage the development and use of similar systems by others who engage in longitudinal studies.

The ACE Cooperative Institutional Research Program

The Cooperative Institutional Research Program is a continuing longitudinal study of students attending a national sample of colleges and universities. The principal purpose of this program of research is to determine how students are affected by different types of college environments. Briefly, the design of the study involves an analysis of differential changes in the interests, achievements, values, and behaviors of students in different types of colleges.

Initial input or "pretest" data are obtained by means of a 150-item questionnaire completed by the incoming freshman during his period of orientation or registration at the college. Output or "posttest" data are obtained through followup questionnaires mailed to the student's home (the student is asked to provide his home address on the initial freshman questionnaire). These followup data are merged with the pretest data to create the longitudinal file which provides the major empirical resource for the research. The current plan for the research program calls for followups after one year, after four years, and at (as yet undetermined) points thereafter. Since new pretest data are obtained from each successive class of entering freshmen, the number of longitudinal research

files increases each year. The current files now include pretest data from more than a million students comprising the entering classes of 1966, 1967, 1968, and 1969 at some 300 institutions. Longitudinal followup data have already been collected from approximately 250,000 of these students (in order to reduce costs, subsamples of students, rather than all of the entering freshmen, are selected for followup at the larger institutions).

When the research program was initiated in 1965 with a pilot study of some 42,000 freshmen at 61 institutions, a more-or-less traditional system of protecting the confidentiality of the data was instituted. The students' responses to the freshman questionnaire were keypunched and converted to magnetic tape. The original questionnaires and punched cards were then destroyed. Following a practice which has frequently been recommended for maintaining longitudinal data files (Dunn, 1967), we created two physically separate tape files. The first file contained the student's answers to the research questions, together with an arbitrary identification number. The second file contained only the student's name and address and the same arbitrary number. Whereas the research data file was openly accessible to members of the ACE Research staff for use in research studies, the name and address file was kept locked in a vault and removed only temporarily when it was necessary to print address labels for followup mailings. Even on these latter occasions, however, the name-and-address file could be released only for brief

periods and only upon written authorization of the ACE Director of Research. Furthermore, the file could not be copied or removed from the data processing center during these periods of temporary release without explicit instructions to this effect from the Director of Research. The formal regulations employed by the data processing center where the name-and-address file was maintained were identical to those outlined in the Department of Defense's Industrial Security Manual (1966).

Some additional security was introduced into the system in 1966, when student questionnaires that could be optically scanned (rather than keypunched) were used. The use of optical scanning eliminates the need for the extensive perusal and handling of documents that is necessitated by keypunching, and minimizes the possibility of improper disclosure of information to data-handling personnel.

It was our impression that this original system offered as much protection as (and, in most cases, more than) other social science research projects against accidental or deliberate extra-legal exploitation of data. We were still concerned, however, that the system did not offer complete protection for the subject against two potential threats to the confidentiality of the data: (1) subpoena by judicial or legislative agencies; and (2) unauthorized disclosure or "snooping" by research staff members who had access to both files.

Development of the "Link" System

The "Link" system of protecting the research data files involves a major elaboration of the original two-file system described above. Debugging of this new system was begun early in 1969, and the system was made fully operational in fall of 1969. Briefly, what we did was to remove the identification numbers from the name-and-address files, and substitute a second, unrelated set of identification numbers. At the same time, we created a third file -- the "Link" file -- which contained only the two sets of numbers: the original numbers from the research data file, and the new numbers from the name-and-address file (note that this Link file represents the only means of linking the subject's identity with his answers to the questions). The final step in establishing the new system was to deposit the Link file at a computer facility located in a foreign country. No copies of the file are kept at ACE or at any other place within the United States.

The nature of the agreement with the foreign computer facility is such that they will neither copy the file nor make it available to outside persons, including research personnel of the American Council on Education. The foreign facility is bound to this agreement even in the event that the American Council on Education should subsequently request that the file be returned. In other words, a basic condition of the agreement is that the foreign facility is under no circumstances to release this Link file to

other individuals or organizations. Thus, both ACE and the foreign agency must violate the agreement before research data can ever again be matched directly with identifying data.

Storing the Link file in a foreign country provides two important protections for the data. One such protection concerns Congressional or judicial subpoena of the files. Since judicial or legislative subpoenas have no validity outside the United States, it would be impossible for Congressional committees or courts to obtain access to information on individual subjects. Thus, even if courts or committees could obtain both the data file and the name-and-address file, there would be no way for them to link up records in one file with records in the other without the Link file. The possibility of using the data files for extralegal harassment of individuals is virtually eliminated also.

A second, perhaps more basic, form of protection concerns possible "snooping" by members of the Research staff. Traditionally, researchers have persuaded subjects to provide them with data under conditions where the guarantee of anonymity is primarily a matter of the researcher's ethics and goodwill. Thus, the possibility of prying or snooping by individual researchers who had access to these "confidential" data almost always existed. The Link system, however, provides protection against even this eventuality. It should be noted that the principle of the Link system does not necessarily involve using a foreign country in order to protect against unwarranted disclosure by individual researchers: The agreement could

as well be between two agencies within the United States. Use of a foreign country, however, does afford the additional protection against subpoena.

Figure 1 shows schematically how the Link system treats questionnaire data provided by freshmen when they first enter college. Questionnaires are first converted to magnetic tape images by means of an optical mark reader. As soon as this conversion has been completed, the questionnaires are destroyed. This conversion process creates three independent files. The first one, shown on the left of Figure 1, contains all of the questionnaire responses provided by the students, in addition to an arbitrary identification number. The second tape file, shown on the far right of Figure 1, contains only the student's name and address, together with a second arbitrary identification number. The Link file shown in the middle contains no data, no name and address, and only the two sets of numbers. This file is stored at a data processing facility in a foreign country. The freshman data file and the name-and-address file are kept at the ACE's Data Processing Center. The name-and-address file, however, is kept locked in a vault and released only long enough to print name and address labels for mailed followups. These followups typically occur during the summer following the student's freshman year, and at the end of his senior year in college. The freshman data file is the only file actually used in research.

The procedures for collecting followup data are diagrammed in

Figure 2. The name-and-address file is released long enough to print name-and-address labels, after which it is replaced in the vault. The labels (which also contain the ID numbers) are applied to the followup questionnaires, which are in turn mailed to the student's home. As soon as the completed questionnaire is returned, it is converted to magnetic tape directly by means of the optical mark reader, after which it is destroyed. Note that, unlike the processing of the freshman questionnaire, no name-and-address file is created; the only information converted to magnetic tape from the followup questionnaire is the student's responses and his ID number. This magnetic tape file is in turn sent to the data processing facility in the foreign country, where it is copied, with the second ID number being replaced by the first. This new file is then sorted on the first ID number, in order to put the records in a different order. This sorted file is then returned to the ACE Office of Research, where the data are merged with the original freshman data provided by the student when he entered college for the first time. This merged data file then is used in the longitudinal research program.

Since the success of the entire longitudinal research program depends on our ability to follow up individual students over time, an additional "backup" copy of the Link file is stored in still another computer facility located in a foreign country. The agreement with this second facility is similar to that with the first: that under no circumstances are they to return the file to us or to

outside facilities other than the first agency. Thus, if the Link file is inadvertently destroyed or otherwise made unusable at the first foreign facility, this facility can in turn request the second facility to send them the backup copy.

An additional protection is afforded by the fact that the optical scanning of the source questionnaires is performed by an independent agency located in a different city from the ACE research office. This agency has been instructed to forward the raw tape images of the followup questionnaires (containing the second identification number) directly to the foreign country facility. The copied tapes (with the first identification number) are sent from the foriegn facility directly to the Office of Research in Washington, D.C. Thus, it is never necessary for the Office of Research to possess a copy of the raw data tape with the second ID number. This fact offers an additional protection to the student in terms of the information he provides on his followup questionnaire; that is, the research staff is not in a position to identify the responses of individual subjects, even in the interim between the initial processing of followup questionnaires and the replacement of the identification numbers. Note that if the document-to-tape processing were done by the Office of Research, it would be possible to identify the responses of individual subjects by linking the name-and-address file with the initial followup file.

An interesting elaboration of this system is that respective educational research agencies located in different countries can

provide such linking services reciprocally. Thus, currently under consideration with one foreign data processing facility is the possibility of an exchange agreement whereby the ACE Office of Research will maintain a link file for the foreign facility and provide similar linking services. Although identification of these specific foreign facilities would not seriously jeopardize the security provided by the Link model, we believe that keeping such foreign facilities anonymous provides some additional protection, particularly against possible theft of the Link file.

Although the Link system may appear at first to be extreme and perhaps unnecessarily expensive and time-consuming, it is no doubt much more economical than most hardware-software computer systems that have been proposed to achieve file security (Weismann, 1967).

The system is consistent with some legal prescriptions for secure data files insofar as it constitutes a set of "mutually insulated data banks" (Schwartz and Orleans, 1967) whose function is to minimize the possibility of disclosure of personal information. To the extent that communications between one data file and another is limited to a code medium, uninterpretable by the agency handling the data, the recommendations of many experts concerning data bank exploitation are also met (e.g., Sawyer and Schecter, 1969; Davidson, 1969).

There are, of course, many alternative models which could be proposed for maintaining confidentiality as a substitute for or as

an augmentation of the strategy proposed in this paper. One such device is specific legislation to provide "privileged" status for social science research data. Legal protection for researchers is unlikely to be adequate by itself, however, since it would not provide the subject with the same kind of protection against the researcher's violation of confidentiality that the Link system does. Misuse of information caused by accidental leakage or by deliberate extralegal exploitation (e.g., commercial usage) is rather difficult to control without well-specified administrative procedures to strengthen the enforcability of legal requirements (Fanwick, 1967; Banshaf, 1968).

Legislation to protect the respondent and researcher may not be feasible or may be slow in enactment. A possible alternative strategy would involve the cooperation of a public agency such as the Census Bureau. Insofar as such an agency can provide Link file services, under legal protection, then the logistical problems associated with the use of foreign facilities can be eliminated. This alternative appears to be a reasonable augmentation of current government concern with protection of research subjects in federally subsidized research projects (U.S. Department of Health, Education and Welfare, 1969).

Consequences of Adopting a Link File System

The implementation of a system for maintaining the security of data and anonymity of respondents has some important social and legal

implications.

One of the major problems confronting the researcher who undertakes any large-scale project is the reluctance of subjects to participate out of a concern that their responses will not remain confidential. These concerns are exacerbated by talk of computerized "dossiers," "national data banks," "invasion of privacy," and the like. The problem here for the researcher is to make it clear to the subject that identifying information is needed not for administrative purposes, but only for updating the file. It seems likely that public understanding of the distinctions among intelligence systems, administrative records, and survey research data would be clarified if the basic concept of the Link system could be adequately communicated. Perhaps the most important educational feature of the system is that it points up the use of identifying information as an accounting device for updating social science research records rather than as a mechanism for evaluating individuals (Astin, 1968).

The effects on research of knowledge of the Link system are testable. Experiments could be designed, for example, to assess the impact of such a system on survey respondents. Any effect could be assessed from differences in response rates or from the precision or accuracy of responses when one survey subsample is provided with information about the Link system and another is not.

For the researcher, adoption of a Link system or similar operation

requires some technical understanding as well as additional time and finances for its development and maintenance. Budget allocations for this purpose subtract from the funds available to support the actual research or analysis. On the positive side, the system does provide a significant increment to the level of protection now afforded most respondents. Any individual researcher must, of course, weigh his concern with maintaining reasonably secure data files against the magnitude of the effort and expense required to implement a Link system or some similar system. Although balancing these objectives may be a difficult task, we feel strongly that the degree of importance of the problem makes the effort worthwhile.

Footnotes

1
The work reported in this paper was supported by grants from the National Science Foundation and the National Institute of Mental Health, and by the general funds of the American Council on Education.

References

Astin, A. W. "Why We Need Your Number." Paper presented at the Annual Meeting of the American Personnel and Guidance Association, Detroit, 1968.

Banshaf, J. F. "When Your Computer Needs a Lawyer." Communications of the ACM, Vol. 2, No. 8, August, 1968.

Boruch, R. F. "Educational Research and the Confidentiality of Data." ACE Research Reports, Vol. 4, No. 4, 1969.

Davidson, T. A. "Computer Information Privacy." The Office, Vol. 70, No. 2, 1969, pp. 10-17.

Department of Defense. Industrial Security Manual for Safeguarding Classified Information. DOD 5220, 22-M. Washington, D. C.: Department of Defense, July 1, 1966.

Dunn, E. S. "The Idea of a National Data Center and the Issue of Personal Privacy." The American Statistician, Vol. 21, No. 1, February, 1967, pp. 21-27.

Fanwick, Charles. "Computer Safeguards: How Safe Are They?" SDC Magazine, Vol. 10, 1967, pp. 26-28.

Nugent, F. A. "Confidentiality in College Counseling Centers." Personnel and Guidance Journal, May, 1969, pp. 872-877.

Privacy and Behavioral Research. Report prepared for Executive Office of the President, Office of Science and Technology. Washington, D. C.: U. S. Government Printing Office, February, 1967.

Sawyer, Jack, and Schechter, Howard. "Computers, Privacy, and the National Data Center: The Responsibility of Social Scientists." American Psychologist, Vol. 23, No. 11, November, 1968.

Swartz, R. D., and Orleans, S. "On Legal Sanctions." University of Chicago Law Review, Vol. 34, 1967, pp. 274-300.

U. S. Department of Health, Education and Welfare. Public Health Service, "Protection of the Individual as a Research Subject--Grants, Awards, Contracts." Washington, D.C.: U.S. Government Printing Office, May, 1969.

Weismann, Clark. "Programming Protection: What Do You Want to Pay?" SDC Magazine, Vol. 10, No. 7, July, 1967, pp. 30-31.

Westin, A. F. Privacy and Freedom. New York: Atheneum, 1967.

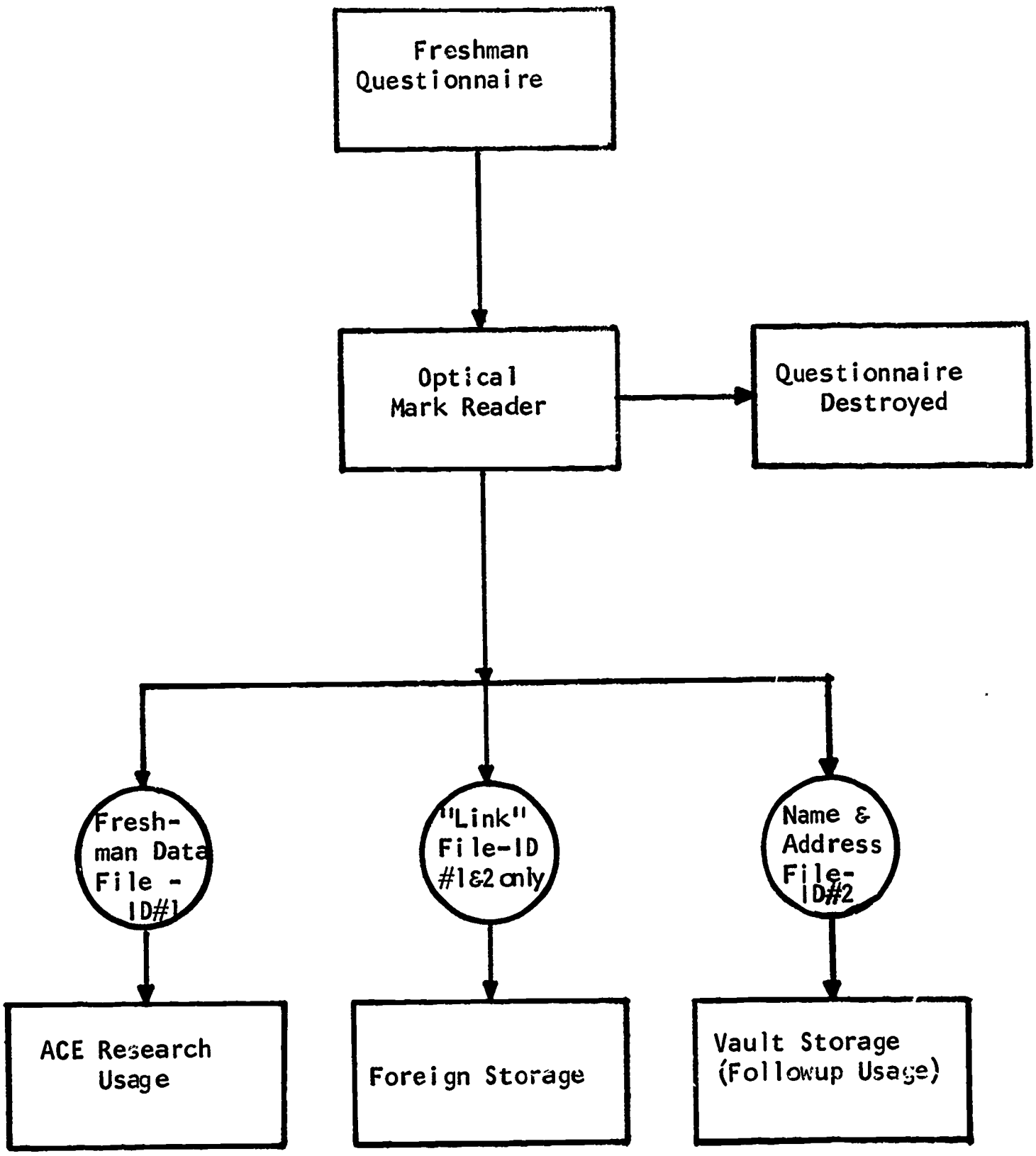

Figure 1: Procedures for Handling Freshman Questionnaire (Pretest) Data.

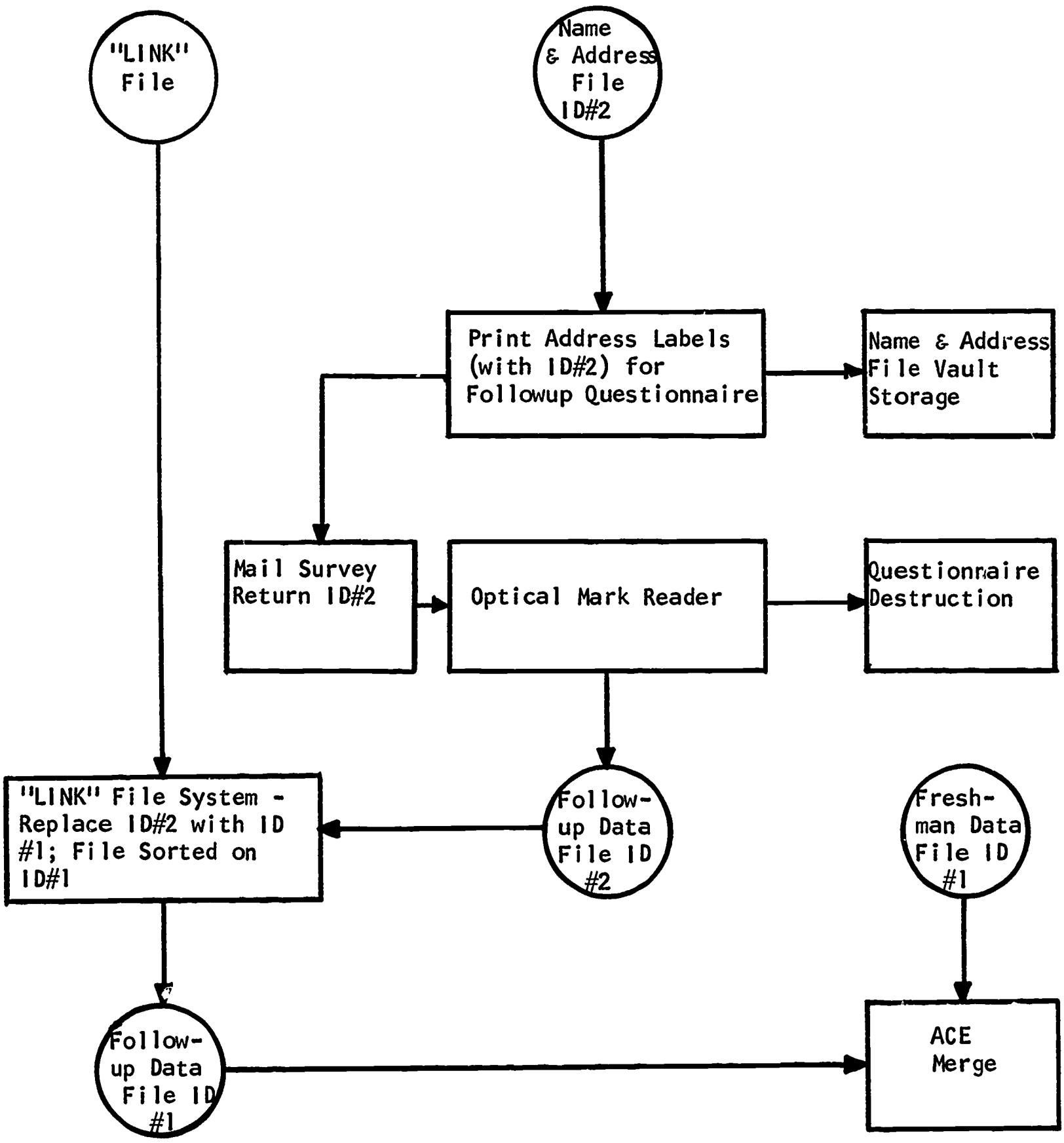

Figure 2: Procedures for Conducting Followup (Post-test) Studies.

Publications by the Staff of the Office of Research
American Council on Education

Astin, A. W. *Recent Findings from the ACE Research Program: Implications for College Choice and Admissions.* **College and University**, Washington: American Association of Collegiate Registrars and Admissions Officers, Summer, 1969. Pp. 341–356.

Astin, A. W. *Racial Considerations in Admissions,* **The Racial Crisis in Higher Education**. Background paper for the Fifty-second Annual Meeting of the American Council on Education. Washington: The Council, 1969. Pp. 65–93.

Astin, A. W. *Folklore of Selectivity,* **The Saturday Review**, December, 1969. Pp. 57–58; 69–70.

Astin, A. W., and Panos, R. J. **The Educational and Vocational Development of College Students.** Washington: The American Council on Education, 1969, 211 pp.

Bayer, A. E. *Marriage Plans and Educational Aspirations,* **American Journal of Sociology**, September, 1969. Pp. 239–244.

Bayer, A. E., and Boruch, R. F. *Black and White Freshmen Entering Four-Year Colleges,* **Educational Record**, Winter, 1969. Pp. 371–386.

Bayer, A. E., and Astin, A. W. *Violence and Disruption on the U.S. Campus: 1968–1969,* **Educational Record**, Winter, 1969. Pp. 337–350.

Bayer, A. E., and Schoenfeldt, L. F. *Student Interchangeability in Three-Year and Four-Year Nursing Programs,* **Journal of Human Resources**, Winter, 1970. Pp. 71–88.

Boruch, R. F. *Faculty Role and Campus Unrest,* **ACE Research Reports**, Vol. 4, No. 5. Washington: American Council on Education, 1969, 28 pp.

Boruch, R. F. *Educational Research and the Confidentiality of Data,* **ACE Research Reports**, Vol. 4, No. 4. Washington: American Council on Education, 1969, 50 pp.

Boruch, R. F. *ACE Research and the Confidentiality Issue,* **Proceedings of the Social Statistics Section, American Statistical Association**, 1969. Pp. 412–417.

Boruch, R. F., and Bayer, A. E. *Financial Resources of Negro College Students: Survey Design and Preliminary Results,* **Proceedings of the Social Statistics Section, American Statistical Association**, 1969. Pp. 389–397.

Creager, J. A. *Fortran Programs Providing Weights in Survey Designs Using Stratified Samples,* **Educational and Psychological Measurement**, Autumn 1969. Pp. 709–712.

Creager, J. A., and Sell, C. L. *The Institutional Domain of Higher Education: A Characteristics File for Research,* **ACE Research Reports**, Vol. 4, No. 6. Washington: American Council on Education, 1969, 83 pp.

Creager, J. A., Astin, A. W., Boruch, R. F., Bayer, A. E., and Drew, D. E. *National Norms for Entering College Freshmen–Fall 1969.* **ACE Research Reports**, Vol. 4, No. 7. Washington: American Council on Education, 1969, 92 pp.

Folger, J. K., Astin, H. S., and Bayer, A. E. **Human Resources and Higher Education.** New York: Russell Sage, 1969, 461 pp.